# Identification of depression and PTSD among Twitter users using pre-trained language model

Anonymous EMNLP submission

## Abstract

Suicide is a global health issue and early diagnosis is necessary for effective treatment. Recent advancements in natural language processing has aided the identification of mental health disorders in social media. This paper investigated the efficacy of pre-trained language model (PLM) in identifying depression and post-traumatic stress disorder (PTSD) with Twitter data. Leveraging the CLPysch 2015 dataset (which constitutes of tweets from users with depression, PTSD and neither condition), we implemented various experimental designs using Long Short Term Memory (LSTM) and attention. The results demonstrate that while performance decreases for multi-nominal classification, the detection of mental health conditions improves with the implementation of attention. This study also underscores the complexity of differentiating between overlapping lexicons with multiple mental health conditions and highlights the potential of PLMs in supporting mental health diagnosis.

## 1 Introduction

Suicide is a global health problem and is the fourth leading cause of death for the 15-44 years demographic globally (World Health Organization, 2021). Mental disorders, including depression and post-traumatic stress disorder (PTSD) have been found to increase the likelihood of suicidal ideation and suicide (Holliday et al., 2021; Busby Grant et al., 2023; Chou et al., 2023; Kratovic et al., 2021). These disorders not only hamper the quality of life for the people who suffer with them but also lessen the quality of life for their families and environment (García-Noguez et al., 2023). Moreover, 75% of people with a severe mental disorder do not receive treatment (Ji et al., 2021). Early diagnosis and subsequent treatment can help to lessen the negative impacts that arise from mental health disorders (Beirão et al., 2020; Kearns et al., 2012).

Researchers are leveraging social context to better understand mental health problems and has been an ongoing process. In the past, researchers used Google trends for mental health surveillance (Page et al., 2011), examining depression based chatter on Twitter (Cavazos-Rehg et al., 2016) and implementing machine learning algorithms to classify tweets in terms of stress or relaxation (Doan et al., 2017). Recently, advancements in natural language processing (NLP) and pre-trained language models (PLMs) have been helpful in identifying the mental health disorder traits from textual data (Ji et al., 2021; Vajre et al., 2021). Although these methods will never fully replace the psychiatric diagnosis and psychotherapy, they assist researchers and clinicians in early detection of mental health symptoms.

Prior to the advancement of PLMs, an early study was conducted in 2014 as a part of a hackathon event (Coppersmith et al., 2014). The authors performed a binary classification between the combinations of control, PTSD and depression outcomes based on the tweets gathered via Twitter api (Coppersmith et al., 2015). Following this research, the same dataset has aided other research, for example, interpreting mental health outcomes (Yang et al., 2023), training new PLMs centric to mental health outcomes (Ji et al., 2021) and comparing various machine learning models for their effectiveness in capturing mental health outcomes (Husseini Orabi et al., 2018).

However, the aforementioned studies focused on binary classification (depression vs control group) to identify the presence or absence of depression among Twitter users. Although there are overlapping expressions between PTSD and depression, there are also dissimilarities between the two mental disorders. Given how these disorders may affect an individual differently, identification of PTSD and depression separately could influence an individual's journey to recovery (Finch, 2023). Proper diagnosis allows clinicians to recommend therapeutic interventions based on specific conditions (Finch, 2023; Kimberly Holland, Timothy J. Legg, 2019). As such, in this research, we extend the classification to all categories of CLPysch 2015 dataset, i.e. depression, PTSD and control, based on tweets.

## 2 Methodology

We aim to answer two key questions in this paper: 1. How effective are PLMs for tracking multiple mental health problems? 2. Which method is most effective for handling multi-nominal mental health classification?
Alongside the two questions, we also scrutinise the scenarios where only depression detection or the detection of general mental health issues might be essential.

### 2.1 CLPysch 2015 shared dataset

The CLPysch 2015 shared dataset contains publicly available tweets collected from the Twitter api over the period 2008 to 2013. The tweets were posted by users with PTSD, depression and a control group who did not have any identified mental health conditions as per tweets (Coppersmith et al., 2014). In total, there are 1145 training set and 599 testing set of anonymous users. Please note that the numbers may not match the original set due to the exclusion of users whose conditions were not recorded.
For this study, we used all available users and their subsequent tweets to identify their category of mental health condition, if present. Since the number of control (572 training, 299 testing) users were higher than depression (327 training, 150 testing) and PTSD (246 training, 150 testing) users, we used weighted cross entropy function for calculation of loss. However, the number of tweets was reduced to a maximum of last 1000 tweets per user out of a possible maximum of 3000 tweets per user due to computational constraints. Despite this, each epoch per experiment took over a day due to the large volume of the dataset and the reliance on the sequential computation of the tweets.

### 2.2 Algorithm for the experimental designs

The experiments were run for all users using Algorithm 1. The number of epochs was set to 20, with the training loop exiting if there was a increase in the training loss. A single user was taken as their own batch for training because of the choice of model designs. Please refer to Section 3 for the model designs. All the tweets went through pre-processing phase where the textual content was cleaned removing any white spaces, retweets, mentions, URLs, punctuation and emoticons. Please note that cross-validation was not feasible due to the magnitude of the dataset.

For each user $u_i$, their individual tweets $t_1, t_2, \ldots, t_n$ were tokenized and passed through a pre-trained RoBERTa model. The details of the choice of PLM is provided in section 2.3. The output was a tensor containing the embedding of the tweet $t_i$. The 768 dimension $[CLS]$ token, which contains the classification information of the entire sentence (Devlin et al., 2018), was extracted for each tweet. For each user, these $[CLS]$ tokens were then stacked to form the tensor of shape $t_{n^{u_i}} \times 768$, where $t_{n^{u_i}}$ was the number of tweets for user $u_i$. Further experiments were performed using these stacked tensors as explained in Section 3. The output of each experiment was then connected to two fully connected layers, with $tanh()$ as the activation function on both layers. The first layer converted the output from 768 dimensions to 100 dimensions and the second layer converted from 100 dimensions to 3 dimensions. The output of the second fully connected layer was passed to softmax function, given by, $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$, to convert the results into probabilities. The final output was the category (control, depression or PTSD) with the highest probability i.e. $max(\sigma(x_i))$.

**Algorithm 1** Training CLPysch 2015 dataset

**for** epochs $(e_i)$ = 1 to $e$ **do**
  **for** users $(u_i)$ = 1 to $u$ **do**
    Pre-process each tweet removing any punctuation, white space, links, retweets and emoticons
    Pass tweet to tokenizer and pre-trained RoBERTa and extract $[CLS]$ token
    Stack all $[CLS]$ tokens for user $u_i$
    Perform experiment $E$ on stacked $[CLS]$ embedding
    Two layers of MLP with $tanh()$ and $softmax()$ to compute predicted $\hat{y}$
    Calculate loss and update weight
  **end for**
  Perform accuracy calculation for epoch $e_i$
**end for**

### 2.3 RoBERTa for base embeddings

We used a Twitter-based fine-tuned model of RoBERTa called *cardiffnlp/twitter-roberta-base* (Barbieri et al., 2020) for the base embeddings as our PLM. The embeddings were extracted using *transformer* library (Wolf et al., 2019). The smaller memory size of RoBERTa and its pre-training on Twitter data made it an appropriate choice for this study. There was an expectancy that the localisation of Twitter vocabulary was present in the PLM of choice. Therefore, it provided appropriate token embeddings for further experiments.

## 3 Experimental Designs

We describe four implemented network models which were used to evaluate the performance of the detection of mental health traits using tweets. The first model used Recurrent Neural Network (RNN), while the remaining three used Attention, which is the engine of transformer-based models. We trained these model on the top of the PLM as described in the section 2.3.

We used a single A100 80GB GPU to train all the models. Each experiment took around 20 days to complete. Hence, the limited number of experiments is due to the lack of resources for performing multiple experiments at once. Please note that the github link containing all the experiments (completed and currently running) will be publicly available in the final paper after the review process.

### 3.1 Long Short Term Memory (LSTM)

In our experiment, we implemented LSTM as our first experiment. Since the tweets are sequential with each user having up to 1000 tweets and there are a differing number of tweets between the users, LSTM was appropriate as an experimental design. Further, LSTM stores long-term dependencies which fails on other neural networks (Hochreiter and Schmidhuber, 1997). We implemented two LSTM models for this research with layers 1 and 2. The added layer increased the complexity of the model. The number of hidden layers in both architectures were set to 100.

### 3.2 Using attention mechanism

Attention is the core of transformer based models (Vaswani et al., 2017). Since we are using RoBERTa for the base model (Barbieri et al., 2020), which is a transformer based model, we added a multi-headed attention layer of 4 heads for our second experiment design. This choice was made to attend to various parts of the tweet sequence differently. The idea behind this design was that the $[CLS]$ token would attend to a single tweet $t_i$ and the stack of $[CLS]$ tokens from each user $t_n^{u_i}$ would use a cross-attention between the tweets i.e. $MHA(t_n^{u_i})$, where $MHA()$ is the multi-head attention. This would determine the presence or absence of some mental health condition (depression or PTSD) for the user $u_i$.

### 3.3 Two sentence sliding window

For this experiment, we used two sentences appended together before the tokenization i.e. for user $u_i$, $t_{u_i} = t_1 + t_2, t_2 + t_3, \ldots, t_{n-1} + t_n$. A sliding window meant that except the first and the last tweet, every tweet in between would have information linked with its previous and the next tweet, creating a short term attention. The resulting $[CLS]$ token would go through cross-attention layer for long term attention across all the tweets belonging to a single user $u_i$, similar to section 3.2.

### 3.4 Adding temporal information

In this experiment, we added temporal information in terms of time lapse between the current and previous tweet as a part of the tweet. The first tweet $t_1$ was converted to $t_1 = "First\ tweet: " + t_1$ and every subsequent tweets were converted to $t_i = "After\ x:," + t_i$, where $x$ was the time lapse between the current tweet $t_i$ and the last tweet $t_{i-1}$, adding temporal context to the tweets. These were then processed in the same fashion as the attention as described in section 3.2.

## 4 Evaluation

We evaluated the experiments based on two key metrics: F1 score and Recall. The best performing results are presented in Table 1. Given $p_1$, $p_2$ and $p_3$ are probabilities for control, depression and PTSD respectively, the results were calculated as such for the mentioned three cases.

**Case A. Multinominal classification:** In this case, we performed the identification of control vs depression vs PTSD users based on the highest probability i.e. $max(p_1, p_2, p_3)$. This was the primary objective of this study.

**Case B. Depression vs Control:** In this case, we removed the probability $p_3$ from all experimental

| Models | Multinominal (A) | | Depression vs Control (B) | | Mental Health vs Control (C) | |
|---|---|---|---|---|---|---|
| | F1 Score | Recall | F1 Score | Recall | F1 Score | Recall |
| LSTM (1 layer) | 0.522 | 0.527 | 0.586 | 0.713 | 0.688 | 0.757 |
| LSTM (2 layers) | 0.504 | 0.524 | 0.569 | **0.827** | 0.712 | **0.913** |
| Attention | 0.595 | 0.590 | 0.616 | 0.780 | 0.724 | 0.830 |
| Temporal | 0.606 | 0.603 | 0.620 | 0.680 | 0.716 | 0.723 |
| 2 sentence Attention | **0.635** | **0.637** | 0.655 | 0.760 | **0.755** | 0.797 |
| MentalRoBERTa | - | - | **0.697** | 0.703 | - | - |

Table 1: Performance metrics across experiments for control vs depression vs PTSD (multinominal) classification (A), depression vs control classification (B) and (depression or PTSD) vs control classfication (C)

results and re-scaled the results for $p_1$ and $p_2$ and evaluated using the readjusted probabilities. This was done to compare our model results with the baseline, MentalRoBERTa (Ji et al., 2021). MentalRoBERTa was taken as the baseline due to its large scale training on mental health texts, including the same data set as ours.

**Case C: Mental health vs Control:** In this case, we added the probability of $p_2$ and $p_3$ from all experimental results and evaluated using the new probability. This was done to simulate a scenario where presence or absence of any mental health condition is tested. This also allowed us to confirm or deny if there are overlapping sentiments among users with depression and PTSD.

In our experiments, two-sentence attention model achieved the best performance in both metrics for case A. Similarly, the same model performed best in F1 score for case C, while recall was higher for 2 layer LSTM for case C. Recall was also higher for 2 layer LSTM in case B. However, for case B, our model did not outperform the baseline F1 score of MentalRoBERTa model.

While our metrics are lower for case A in comparison to other cases, it is expected of a multinominal classification compared to binary classification. Identification of depression and PTSD separately resulted in decreased performance, compared to case B where only depression is identified and case C where general mental health condition is identified. Another possible explanation is the potential overlap of expressions in tweets from users with depression and PTSD. Consequently, the classification between the two groups becomes more challenging compared to the classification of an individual mental disorder from the control group alone. However, when these disorders are combined, the result improves significantly as seen from the results in case C of Table 1.

High values of recall in 2 layer LSTM for case B (0.827) and case C (0.913) also means that majority of mental health users are identified. While this causes less generalisation as demonstrated by their corresponding F1 score, it is desirable for this particular study because not identifying mental health users are more costly than identifying false positives of the same.

It should be noted that building state-of-the-art model was not the primary objective of this study. Instead, it was a study to target identification of multiple mental health disorder for early diagnosis. Further, these models cannot replace psychiatric diagnosis and therapeutic interventions, but they are valuable tools to aid clinicians and researchers.

## 5 Conclusion

In this study, we implemented and evaluated PLM efficacy in identifying multiple mental health conditions, including depression and PTSD from Twitter data. Our experiments, including LSTM, attention, sliding window approach, and the integration of temporal information, showed that the two-sentence attention model performs adequately for detecting multiple health conditions. While the performance was not as high as binary identification, it can be attributed due to the overlap of sentiments in tweets between depression and PTSD users. Our findings also indicate that two layer LSTM model is better at detecting the presence of depression or mental health, in general, but it failed to generalise well. In this regard, perhaps attention based mechanism was significant as well. Despite lower metrics in multi-nominal setting, our study provides an avenue of early mental health detection, potentially leading to better targeted treatment and interventions using social media.

## Limitations

One of the aforementioned limitations is that only last 1000 tweets (if more than 1000 tweets present) per user were considered for this research. The GPU server was shared between various projects as well as the lack of resources to add more GPU servers meant that not all tweets could be processed. The reliance on the processing of tweets sequentially further meant that each epoch was much longer, since batching was not possible. This caused each model to run up to 20 days, hence resulting in lower number of experiments. Further, only a single dataset was used, which could bias the results. In addition, the tweets were extracted a decade ago, which means the newer tweets would not have been collected. The lexicon in which humans express sentiments perhaps changed in the last decade and those were not captured. Additionally, the collected tweets are only a sub-sample of the much larger cohort of mental health users who are not considered in this study. Even while focusing on this cohort itself, there is a lack of evidence to affirm the presence or absence of mental health conditions between the Twitter users. Finally, our study aims to develop a model for assisting researchers and clinicians for detection of mental health conditions using social context for non-clinical use. However, it does not replace clinical diagnoses which is essential for the detection and treatment of mental health issues.

## Ethics Statement

The ethics was approved in accordance to Human Research Ethics Committee (HREC) approval number H15559. The data was already de-identified when it was received from Department of Computer Science, John Hopkins University.


## Acknowledgements

We would like to thank Professor Mark Dredze for providing data and Department of Computer, Data and Mathematical Sciences, Western Sydney University for allowing to use the GPU clusters which allowed for processing of data.


## References


Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. ArXiv:2010.12421 [cs].

Diogo Beirão, Helena Monte, Marta Amaral, Alice Longras, Carla Matos, and Francisca Villas-Boas. 2020. Depression in adolescence: a review. *Middle East Current Psychiatry*, 27(1):50.

Janie Busby Grant, Philip J. Batterham, Sonia M. McCallum, Aliza Werner-Seidler, and Alison L. Calear. 2023. Specific anxiety and depression symptoms are risk factors for the onset of suicidal ideation and suicide attempts in youth. *Journal of Affective Disorders*, 327:299–305.

Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut. 2016. A content analysis of depression-related tweets. *Computers in Human Behavior*, 54:351–357.

Po-Han Chou, Shao-Cheng Wang, Chi-Shin Wu, and Masaya Ito. 2023. Trauma-related guilt as a mediator between post-traumatic stress disorder and suicidal ideation. *Frontiers in Psychiatry*, 14. Publisher: Frontiers.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Type: Journal article.

Son Doan, Amanda Ritchart, Nicholas Perry, Juan D Chaparro, and Mike Conway. 2017. How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets. *JMIR Public Health and Surveillance*, 3(2):e35.

Jon Finch. 2023. The Difference Between Depression and PTSD.

Luis Roberto García-Noguez, Saúl Tovar-Arriaga, Wilfrido Jacobo Paredes-García, Juan Manuel Ramos-Arreguín, and Marco Antonio Aceves-Fernandez. 2023. Automatic classification of depressive users on Twitter including temporal analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1):38.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. Type: Journal article.

Ryan Holliday, Claire A. Hoffmire, W. Blake Martin, Rani A. Hoff, and Lindsey L. Monteith. 2021. Associations between justice involvement and PTSD and depressive symptoms, suicidal ideation, and suicide attempt among post-9/11 veterans. *Psychological Trauma: Theory, Research, Practice, and Policy*, 13(7):730–739. Place: US Publisher: Educational Publishing Foundation.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. ArXiv:2110.15621 [cs].

Megan C. Kearns, Kerry J. Ressler, Doug Zatzick, and Barbara Olasov Rothbaum. 2012. Early Interventions for Ptsd: A Review. *Depression and Anxiety*, 29(10):833–842. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.21997.

Kimberly Holland, Timothy J. Legg. 2019. PTSD and Depression: Similarities, Differences & What If You Have Both.

Layla Kratovic, Lia J. Smith, and Anka A. Vujanovic. 2021. PTSD Symptoms, Suicidal Ideation, and Suicide Risk in University Students: The Role of Distress Tolerance. *Journal of Aggression, Maltreatment & Trauma*, 30(1):82–100. Publisher: Routledge _eprint: https://doi.org/10.1080/10926771.2019.1709594.

Andrew Page, Shu-Sen Chang, and David Gunnell. 2011. Surveillance of Australian Suicidal Behaviour Using the Internet? *Australian & New Zealand Journal of Psychiatry*, 45(12):1020–1022. Publisher: SAGE Publications Ltd.

Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30. Type: Journal article.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

World Health Organization. 2021. Suicide worldwide in 2019: global health estimates. *World Health Organization*. Type: Journal article.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models.